Overview of the ACLIA IR4QA (Information Retrieval for Question Answering) Task

Tetsuya Sakai  Noriko Kando
Chuan-Jie Lin  Teruko Mitamura
Donghong Ji    Kuang-Hua Chen
Eric Nyberg

18th December 2008 @NTCIR-7, Tokyo
TALK OUTLINE

1. Task Objectives
2. Relevance Assessments
3. Evaluation Metrics
4. Participating Teams
5. Official Results
6. Lazy Evaluation
7. Unanswered Questions
What are the effective IR techniques for QA?

Notes:
1. CLIR system can take a natural language question in source language.
2. In collaboration with CLQA, CLIR system can also take translated keyterms and answer type analysis.
3. Translation often happens in here.
Traditional “ad hoc” IR vs IR4QA

• Ad hoc IR (evaluated using Average Precision etc.)
  - Find as many (partially or marginally) relevant documents as possible and put them near the top of the ranked list

• IR4QA (evaluating using… WHAT? )
  - Find relevant documents containing different correct answers?
  - Find multiple documents supporting the same correct answer to enhance reliability of that answer?
  - Combine partially relevant documents A and B to deduce a correct answer?
TALK OUTLINE

1. Task Objectives
2. Relevance Assessments
3. Evaluation Metrics
4. Participating Teams
5. Official Results
6. Lazy Evaluation
7. Unanswered Questions
Pooling for relevance assessments

Target Documents

- CS: Simplified Chinese
- CT: Traditional Chinese
- JA: Japanese

Relevance assessments:

- L2: relevant
- L1: partially relevant
- L0: judged nonrelevant
Different pool depths for different topics

Assess depth-30 pool

Assess depth-50 pool (minus depth-30 pool)

Assess depth-70 pool (minus depth-50 pool)

Assess depth-90 pool (minus depth-70 pool)

Assess depth-100 pool (minus depth-90 pool)

Mandatory for all topics

See IR4QA Overview Tables 29-31 for details

Relevance assessments coordinated independently by Donghong Ji (CS), Chuan-Jie Lin (CT) and Noriko Kando (JA)
Sorting the pooled documents for assessors

- Traditional approach: Docs sorted by IDs
- IR4QA approach: Sort docs in depth-X pool by:
  - #runs containing the doc at or above rank X (primary sort key)
  - Sum of ranks of the doc within these runs (secondary sort key)

Present ``popular” documents first!
Assumptions behind the sort

1. Popular docs are more likely to be relevant than others.

Supported by [Sakai and Kando EVIA 08]

2. If relevant docs are concentrated near the top of the list to be assessed, this is easier for the assessors to judge more efficiently and consistently.

At NTCIR-2, the assessors actually did not like doc lists sorted by doc IDs

(But we need more empirical evidence)
TALK OUTLINE

1. Task Objectives
2. Relevance Assessments
3. Evaluation Metrics
4. Participating Teams
5. Official Results
6. Lazy Evaluation
7. Unanswered Questions
Average Precision (AP)

\[ AP = \frac{1}{R} \sum_r I(r) \left( \frac{C(r)}{r} \right) \]

- Number of relevant docs
- 1 iff doc at rank r is relevant
- Precision at rank r

• Used widely since the advent of TREC
• Mean over topics is referred to as “MAP”
• Cannot handle graded relevance (but many IR researchers just love it)
Q-measure (Q)

\[ Q\text{-measure} = \frac{1}{R} \sum_{r} I(r) \frac{C(r) + \beta cg(r)}{r + \beta cg^*(r)} \]

- Generalises AP and handles graded relevance
- Properties similar to AP and higher discriminative power
- Not widely-used, but has been used for QA and INEX as well as IR

Persistence Parameter $\beta$ set to 1

Blended ratio at rank $r$
(Combines Precision and normalised Cumulative Gain)

Sakai and Robertson EVIA 08 provides a user model for AP and Q
nDCG (Microsoft version)

$$nDCG = \frac{\sum_{r=1}^{l} g(r) / \log(r + 1)}{\sum_{r=1}^{l} g^*(r) / \log(r + 1)}$$

- Fixes a bug of the original nDCG
- But lacks a parameter that reflects the user’s persistence
- Most popular graded-relevance metric

Sum of discounted gains for a system output

Sum of discounted gains for an *ideal* output
IR4QA evaluation package
(Works for ad hoc IR in general)

Computes AP, Q, nDCG, RBP, NCU
[Sakai and Robertson EVIA 08] and so on

http://research.nii.ac.jp/ntcir/tools/ir4qa_eval-en
TALK OUTLINE

1. Task Objectives
2. Relevance Assessments
3. Evaluation Metrics
4. Participating Teams
5. Official Results
6. Lazy Evaluation
7. Unanswered Questions
**Table 1. IR4QA participants.**

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRKLY</td>
<td>University of California, Berkeley</td>
</tr>
<tr>
<td>CMUJAV</td>
<td>Language Technologies Institute, Carnegie Mellon University</td>
</tr>
<tr>
<td>CYUT</td>
<td>Chaoyang University of Technology</td>
</tr>
<tr>
<td>HIT</td>
<td>Heilongjiang Institute of Technology User Group: HIT2 NLP Joint Lab</td>
</tr>
<tr>
<td>KECIR</td>
<td>Shenyang Institute of Aeronautical Engineering</td>
</tr>
<tr>
<td>MITEI</td>
<td>Institute of Computing Technology, Chinese Academy of Sciences</td>
</tr>
<tr>
<td>NLPAI</td>
<td>College of Computer Science and Technology, Wuhan University of Science and Technology</td>
</tr>
<tr>
<td>NTUBROWS</td>
<td>CSIE, National Taiwan University</td>
</tr>
<tr>
<td>OT</td>
<td>Open Text Corporation</td>
</tr>
<tr>
<td>RALI</td>
<td>University of Montreal</td>
</tr>
<tr>
<td>TA</td>
<td>Toyohashi University of Technology</td>
</tr>
<tr>
<td>WHUCC</td>
<td>Computer Center of Wuhan University</td>
</tr>
</tbody>
</table>

- 12 participants from China/Taiwan, USA, Japan
- 40 CS runs (22 CS-CS, 18 EN-CS)
- 26 CT runs (19 CT-CT, 7 EN-CT)
- 25 JA runs (14 JA-JA, 11 EN-JA)

**Monolingual**

**Crosslingual**
Oral presentations

• RALI (CS-CS, EN-CS, CT-CT, EN-CT)
  - Uses Wikipedia to extracts cue words for BIOGRAPHY; Extracts person names using Wikipedia and Google; Uses Google translation
• CYUT (EN-CS, EN-CT, EN-JA)
  - Uses Wikipedia for query expansion and translation; Uses Google translation
• MITEL (EN-CS, CT-CT)
  - Uses SMT and Baidu for translation; data fusion
• CMUJAV (CS-CS, EN-CS, JA-JA, EN-JA)
  - Proposes Pseudo Relevance Feedback using Lexico-Semantic Patterns (LSP-PRF)
Other interesting approaches

- BRKLY (JA-JA) A very experienced TREC/NTCIR participant
- HIT (EN-CS) PRF most successful
- KECIR (CS-CS) Query expansion length optimised for each question type (definition, biography…)
- NLPAI (CS-CS) Uses question analyses files from other teams (next slide)
- NTUBROWS (CT-CT) Query term filtering, data fusion
- OT (CS-CS, CT-CT, JA-JA) Data fusion-like PRF
- TA (EN-JA) SMT document translation from NTCIR-6
- WHUCC (CS-CS) Document reranking

Please visit the posters of all 12 IR4QA teams!
NLPAI (CS-CS) used question analysis files from other teams.

CSWHU-CS-CS-01-T:
  <KEYTERMS>
  <KEYTERM SCORE="1.0">宇宙大爆炸</KEYTERM>
  <KEYTERM SCORE="0.3">理论</KEYTERM>
  </KEYTERMS>

Apath-CS-CS-01-T:
  <KEYTERMS>
  <KEYTERM SCORE="1.0">宇宙大爆炸理论</KEYTERM>
  </KEYTERMS>

CMUJAV-CS-CS-01-T:
  <KEYTERMS>
  <KEYTERM SCORE="1.0">宇宙</KEYTERM>
  <KEYTERM SCORE="1.0">爆炸</KEYTERM>
  <KEYTERM SCORE="1.0">理论</KEYTERM>
  <KEYTERM SCORE="1.0">宇宙大爆炸理论</KEYTERM>
  <KEYTERM SCORE="1.0">宇宙大爆炸</KEYTERM>
  </KEYTERMS>

Different teams come up with different set of query terms with different weights. This clearly affects retrieval performance.

Special thanks to Maofu Liu (NLPAI)
TALK OUTLINE

1. Task Objectives
2. Relevance Assessments
3. Evaluation Metrics
4. Participating Teams
5. Official Results
6. Lazy Evaluation
7. Unanswered Questions
CS T-runs: Top 3 teams

<table>
<thead>
<tr>
<th></th>
<th>Mean AP</th>
<th>Mean Q</th>
<th>Mean nDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>OT-CS-CS-04-T</td>
<td>.6337</td>
<td>.6490</td>
<td>.8270*</td>
</tr>
<tr>
<td>MITEL-EN-CS-03-T</td>
<td>.5959</td>
<td>.6124</td>
<td>.7951</td>
</tr>
<tr>
<td>CMUJAV-CS-CS-02-T</td>
<td>.5930</td>
<td>.6055</td>
<td>.7949</td>
</tr>
</tbody>
</table>

- MITEL is very good even though it is a crosslingual run
- OT significantly outperforms CMUJAV with Mean nDCG (two-sided bootstrap test; \( \alpha=0.05 \))
- nDCG disagrees with AP and Q
## CT T-runs: Top 3 teams

<table>
<thead>
<tr>
<th></th>
<th>Mean AP</th>
<th>Mean Q</th>
<th>Mean nDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>MITEL-CT-CT-02-T</td>
<td>.5839</td>
<td>.6018</td>
<td>.7873</td>
</tr>
<tr>
<td>OT-CT-CT-04-T</td>
<td>.5521**</td>
<td>.5724**</td>
<td>.7656**</td>
</tr>
<tr>
<td>RALI-CT-CT-05-T</td>
<td>.3952</td>
<td>.4096</td>
<td>.6559**</td>
</tr>
</tbody>
</table>

- MITEL and OT not significantly different from each other
- OT significantly outperforms RALI (two-sided bootstrap test; $\alpha=0.01$) but RALI’s performance is actually very high after bug fix
## JA T-runs: Top 3 teams

<table>
<thead>
<tr>
<th>Mean AP</th>
<th>Mean Q</th>
<th>Mean nDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>OT-JA-JA-04-T</td>
<td>OT-JA-JA-04-T</td>
<td>OT-JA-JA-04-T</td>
</tr>
<tr>
<td>.6979 **</td>
<td>.7090 **</td>
<td>.8650 **</td>
</tr>
<tr>
<td>CMUJAV-JA-JA-01-T</td>
<td>CMUJAV-JA-JA-01-T</td>
<td>CMUJAV-JA-JA-01-T</td>
</tr>
<tr>
<td>.5932</td>
<td>.5996</td>
<td>.7832</td>
</tr>
<tr>
<td>BRKLY-JA-JA-02-T</td>
<td>BRKLY-JA-JA-02-T</td>
<td>BRKLY-JA-JA-02-T</td>
</tr>
<tr>
<td>.5838 **</td>
<td>.5996 **</td>
<td>.7831 **</td>
</tr>
</tbody>
</table>

- OT significantly outperforms CMUJAV
- BRKLY significantly outperforms the 4th team (CYUT crosslingual run)

(two-sided bootstrap test; α=0.01)
System ranking by Q/nDCG vs that by AP

By definition, nDCG is more forgiving for low-recall runs than AP and Q.
The most “novel” runs

**RALI-EN-CS-04-T** found 63 unique relevant docs (53 for topic CS-T42)
**RALI-EN-CT-05-T** found 32 unique relevant docs (16 for topic CT-T442)
**OT-JA-JA-01-T** found 51 unique relevant docs (12 for JA-T236)

These runs are valuable for making the relevance assessments as exhaustive as possible.
## Successful PRF

<table>
<thead>
<tr>
<th></th>
<th>Mean AP</th>
<th>Mean Q</th>
<th>Mean nDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIT-EN-CS-01-DN</td>
<td>.5690**</td>
<td>.5840 **</td>
<td>.7560 **</td>
</tr>
<tr>
<td>HIT-EN-CS-02-DN</td>
<td>.4634</td>
<td>.4827</td>
<td>.6910</td>
</tr>
<tr>
<td>OT-CT-CT-04-T</td>
<td>.5521 **</td>
<td>.5724 **</td>
<td>.7656 **</td>
</tr>
<tr>
<td>OT-CT-CT-02-T</td>
<td>.5111</td>
<td>.5339</td>
<td>.7432</td>
</tr>
<tr>
<td>BRKLY-JA-JA-02-T</td>
<td>.5838 *</td>
<td>.5996 **</td>
<td>.7831 **</td>
</tr>
<tr>
<td>BRKLY-JA-JA-03-T</td>
<td>.5407</td>
<td>.5509</td>
<td>.7475</td>
</tr>
<tr>
<td>OT-JA-JA-04-T</td>
<td>.6979 *</td>
<td>.7090 *</td>
<td>.8650 **</td>
</tr>
<tr>
<td>OT-JA-JA-02-T</td>
<td>.6698</td>
<td>.6808</td>
<td>.8473</td>
</tr>
</tbody>
</table>

Other teams appear to be less successful with PRF. This may be partly because the qrels are very incomplete.
Per-topic AP/Q/nDCG averaged over runs (CS)

“Topic difficulty” varies
Per-topic AP/Q/nDCG averaged over runs (CT)

“Topic difficulty” varies
Per-topic AP/Q/nDCG averaged over runs (JA)

“Topic difficulty” varies
TALK OUTLINE

1. Task Objectives
2. Relevance Assessments
3. Evaluation Metrics
4. Participating Teams
5. Official Results
6. Lazy Evaluation
7. Unanswered Questions
Forming pseudo-qrels

QUESTION: Can we get away with not doing any relevance assessments at all?

1. Sort pooled docs by
   (1) Number of runs that retrieved it; and then
   (2) Sum of its ranks within these runs.

2. Take the top 10 docs in the sorted pool and treat them all as L1-relevant!

Sakai and Kando EVIA 08 actually shows that the top 10 docs are more likely to be relevant than others on average
System ranking by real MAP vs that by pseudo MAP (CS)

“Pseudo MAP” assumes that “popular” documents are relevant.
System ranking by real MAP vs that by pseudo MAP (CT)
System ranking by real MAP vs that by pseudo MAP (JA)

Pseudo-qrels are not very useful for predicting the ranking of the highest performers.

Kendall’s rank correlation
Pseudo vs real: around 0.7 (cf. Soboroff SIGIR 01: around 0.4)

But they may be useful for predicting the low performers (for CT and JA)
TALK OUTLINE

1. Task Objectives
2. Relevance Assessments
3. Evaluation Metrics
4. Participating Teams
5. Official Results
6. Lazy Evaluation
7. Unanswered Questions
Unanswered Questions

• What IR strategies are good for QA? (e.g. How does question classification help?)
• What are the general/language-specific challenges for mono/crosslingual IR4QA?
• How incomplete are the IR4QA test collections? How reusable are they?
• What are the best evaluation methods?
• How do IR4QA and the entire ACLIA results correlate?